Using the Segmentation Corpus to define an inventory of concatenative units for Cantonese speech synthesis

Wai Yi Peggy WONG
Chris BREW
Mary E. BECKMAN
Linguistics Dept., Ohio State University
222 Oxley Hall, 1712 Neil Ave.
Columbus, OH, 43210-1298 USA
{pwong, cbrew, mbeckman}@ling.osu.edu

Shui-duen CHAN
Chinese Language Centre,
Hong Kong Polytechnic University
Yuk Choi Road, Hung Hom, Kowloon,
HONG KONG
chsdchan@polyu.edu.hk

Abstract

The problem of word segmentation affects all aspects of Chinese language processing, including the development of text-to-speech synthesis systems. In synthesizing a Hong Kong Cantonese text, for example, words must be identified in order to model fusion of coda [p] with initial [h], and other similar effects that differentiate word-internal syllable boundaries from syllable edges that begin or end words. Accurate segmentation is necessary also for developing any list of words large enough to identify the word-internal cross-syllable sequences that must be recorded to model such effects using concatenated synthesis units. This paper describes our use of the Segmentation Corpus to constrain such units.

Introduction

What are the best units to use in building a fixed inventory of concatenative units for an unlimited vocabulary text-to-speech (TTS) synthesis system for a language? Given a particular choice of unit type, how large is the inventory of such units for the language, and what is the best way to design materials to cover all or most of these units in one recording session? Are there effects such as prosodically conditioned allophony that cannot be modeled well by the basic unit type? These are questions that can only be answered language by language, and answering them for Cantonese[1] poses several interesting challenges.

One major challenge involves the definition of the "word" in Cantonese. As in other varieties of Chinese, morphemes in Cantonese are typically monosyllabic and syllable structure is extremely simple, which might suggest the demi-syllable or even the syllable (Chu & Ching, 1997) as an obvious basic unit. At the same time, however, there are segmental "sandhi" effects that conjoin syllables within a word. For example, when the morpheme 集 zaap6[2] stands as a word alone (meaning 'to collect'), the [p] is a glottalized and unreleased coda stop, but when the morpheme occurs in the longer word 集合 zaap6hap6 ('to assemble'), the coda [p] often resyllabifies and fuses with the following [h] to make an initial aspirated stop. Accurate word segmentation at the text analysis level is essential for identifying the domain of such sandhi effects in any full-fledged TTS system, whatever method is used for generating the waveform from the specified pronunciation of the word. A further challenge is to find a way to capture such sandhi effects in systems that use concatenative methods for waveform generation.

This paper reports on research aimed at defining an inventory of concatenative units for Cantonese using the Segmentation Corpus, a lexicon of 33k words extracted from a large corpus of Cantonese newspaper texts. The corpus is described further in Section 2 after an excursus (in Section 1) on the problems posed

[1] We use "Cantonese" to mean the newer Hong Kong standard, and not the older Canton City one.

[2] We use the Jyutping romanization developed by the Linguistics Society of Hong Kong in 1993. See http://www.cpct92.cityu.edu.hk/lshk.

by the Cantonese writing system. Section 3 outlines facts about Cantonese phonology relevant to choosing the concatenative unit, and Section 4 calculates the number of units that would be necessary to cover all theoretically possible syllables and sequences of syllables. The calculation is done for three models: (1) syllables, as in Chu & Ching (1997), (2) Law & Lee's (2000) mixed model of onsets, rhymes, and cross-syllabic rhyme-onset units, and (3) a positionally sensitive diphone model. This section closes by reporting how the number of units in the last model is reduced by exploiting the sporadic and systematic phonotactic gaps discovered by looking for words exemplifying each possible unit in the Segmentation Corpus.

1 The Cantonese writing system

The Cantonese writing system poses unique challenges for developing online lexicons, not all of which are related to the "foremost problem" of word segmentation. These problems stem from the long and rich socio-political history of the dialect, which makes the writing system even less regular than the Mandarin one, even though Cantonese is written primarily with the same logographs ("Chinese characters").

The main problem is that each character has several readings, and the reading cannot always be determined based on the immediate context of the other characters in a multisyllabic word. For some orthographic forms, the variation is stylistic. For example, the word 支援 'support' can be pronounced zi1jyun4 or zi1wun4. But for other orthographic forms, the variation in pronunciation corresponds to different words, with different meanings. For example, the character sequence 正當 writes both the function word zing3dong1 'while' and the content word zing3dong3 'proper'. Moreover, some words, such as ko1 'to page, telephone', ge3 (genitive particle), and laa3 (aspect marker), have no standard written form. In colloquial styles of writing, these forms may be rendered in non-standard ways, such as using the English source word *call* to write ko1, or writing the particles with special characters unique to Cantonese. In more formal writing, however, such forms must be left to the reader to interpolate from a character "borrowed" from some other morpheme. For example, ge3 (genitive particle) might be written 的, a character which more typically writes the morpheme dik1 in muk6dik1 目的 'aim', but which suggests ge3 because it also writes a genitive particle in Mandarin (de in the Pinyin romanization). Thus, 的 has a reading dik1 that is etymologically related to the Mandarin morpheme, but it also has the etymologically independent reading ge3 because Cantonese readers can read texts written in Mandarin as if they were Cantonese. Such ambiguities of reading make the task of developing online wordlists from text corpora doubly difficult, since word segmentation is only half the task.

2 The Segmentation Corpus

The Segmentation Corpus is an electronic database of around 33,000 Cantonese word types extracted from a 1.7 million character corpus of Hong Kong newspapers, along with a tokenized record of the text. It is described in more detail in Chan & Tang (1999). The Cantonese corpus is part of a larger database of segmented Chinese texts, including Mandarin newspapers from both the PRC and Taiwan. The three databases were created using word-segmentation criteria developed by researchers at the Chinese Language Centre and Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University. Since these criteria were intended to be applicable to texts in all three varieties, they do not refer to the phonology.

For our current purpose, the most useful part of the Segmentation Corpus is the wordlist proper, a file containing a separate entry for each word type identified by the segmentation criteria. Each entry has three fields: the orthographic form(s), the pronunciation(s) in Jyutping, and the token frequency in the segmented newspaper corpus. In the original corpus, the first field could have multiple entries. For example, there are two character strings, 泛濫 and 氾濫, in the entry for the word faan6laam6 'to flood'. However, the two readings of 支援 were not listed separately in the pronunciation field for that orthographic form (and there was only one entry for the two words written with 正當).

Before we could use the wordlist, therefore, we had to check the pronunciation field for each entry. The first author, a native speaker of Hong Kong Cantonese, examined each entry in order to add variant readings not originally listed (as in the entry for 支援 'support') and to correct readings that did not correspond to the modern Hong Kong pronunciation (as in the entry for 盒 'box'). In addition, when the added variant pronunciation corresponded to an identifiably different word (as in zing3dong3 'proper' versus zing3dong1 'while' for 正當), the entry was split in two, and all tokens of that character string in the larger corpus were examined, in order to allocate the total token count for the orthographic form to the two separate frequencies for the two different words. Approximately 90 original entries were split into separate entries by this processing. In this way, the 32,840 entries in the original word list became 33,037 entries. Once this task was completed, we could use the wordlist to count all of the distinct units that would be needed to synthesize all of the words in the Segmentation Corpus. To explain what these units are, we need to describe the phonology of Cantonese words and longer utterances.

3 Cantonese phonology

The smallest possible word is a nasal ([m] or [ŋ]) or vowel as a bare tone-bearing syllable nucleus, as in 五 ng5 'five' and 啞 aa2 'dumb'. A syllable with a vowel as nucleus can also have an onset consonant, and it can have a more complex rhyme which combines the vowel with any of the three coda nasals [m], [n], and [ŋ], the two glides [j] and [w], or the three coda stops [p], [t], and [k], as in 將 zoeng1 'will', 胃 wai6 'stomach', and 率 leot2 'rate'. Tables 1 and 2 show the 11 vowels and 19 consonants of Cantonese, and Figure 1 plots representative F0 contours for each of the tones that contrast on a syllable with an all-sonorant rhyme.

If there were no phonotactic restrictions on combinations of vowels and consonants, there would be 101 distinct rhymes in each of the six tones. However, not all combinations are possible. For example, tone 5 does not occur on syllables with coda stops (and tone 4 on such checked syllables is limited to onomatopoeia). Also, the mid short vowel [e] does not occur in open syllables, and in closed syllables it occurs only before [k], [ŋ], and [j], whereas [i:] occurs in closed syllables only before [p], [t], [m], [n], and [w]. The Jyutping transliteration system takes advantage of this kind of complementary distribution to limit the number of vowel symbols. Thus "i" is used to transcribe both the short mid vowel [e] in the rhymes [ek] and [eŋ], and the high vowel [i:] in the rhymes [i:], [i:t], [i:p], [i:m], [i:n], and [i:w]. Ignoring tone, there are 54 rhyme types in standard usage of Jyutping. Canonical forms of longer words can be described to a very rough first approximation as strings of syllables. However, fluent synthesis cannot be achieved simply by stringing syllables together without taking into account the effects of position in the word or phrase.

front		central	back	
	round		round	
i:	y:		u:	high
e		ɵ	o	mid (short)
ɛ:	œ:		ɔ:	mid (long)
		ɐ	a:	low vowels

Table 1. Vowels of Cantonese.

labial	dental	palatal	velar	labio velar	
p^h	t^h, ts^h		k^h	k^{hw}	
p	t, ts		k	k^w	
f	s				h
m	n, l	j	ŋ	w	

Table 2. Consonants of Cantonese.

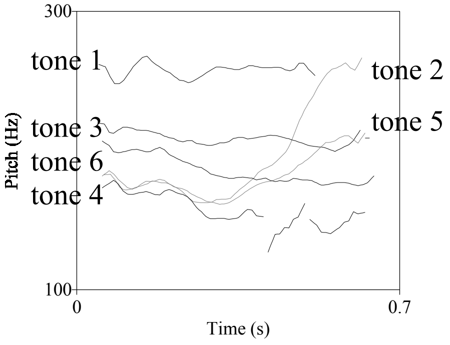

Figure 1. F0 contours for six words [wɐj] with different tones. Numbers to the right identify the endpoints of the two rising tones (in grey) and numbers to the left identify starting points of the other four tones (in black). The discontinuities in wai4 are where the speaker breaks into creak.

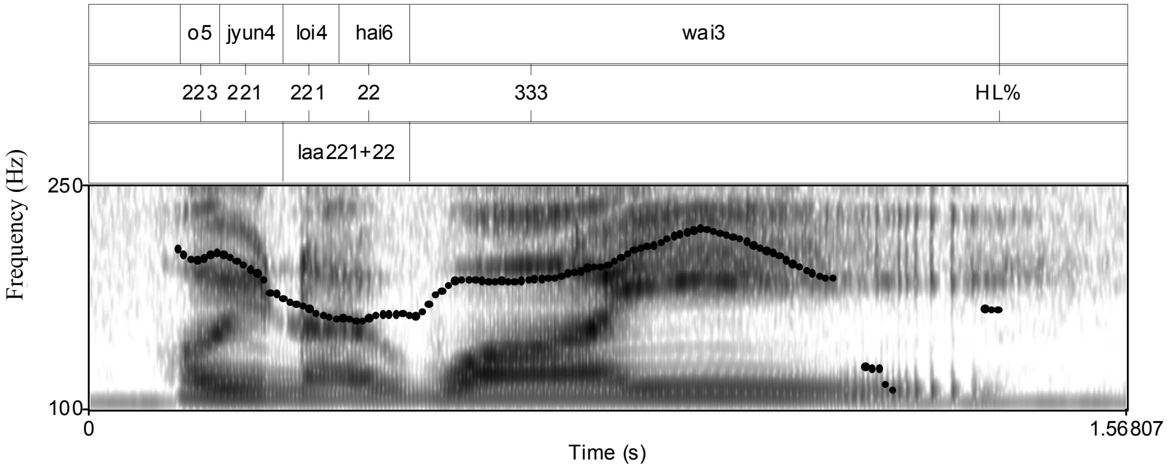

Figure 2. Spectrogram with F0 track superimposed for an utterance of the sentence `o5 jyun4loi4 hai6 wai3` 'Oh, I get it. It was the character 慰!' (The context is a dictation task.) The labelling window above the signal view shows a partial transcription in the annotation conventions proposed by Wong, Chan & Beckman (in press), with a syllable-by-syllable Jyutping transliteration (top tier), a transcription of the (canonical) lexical tones and boundary tone, and a phonetic transcription of fused forms (lowest tier). The HL% boundary tone is a pragmatic morpheme, which we have translated with the 'Oh, I get it.' phrase.

		經		濟			學	家		basic units
Jyutping		ging1		zai3			hok6	gaa1		(added units)
Chu & Ching	keŋ			tsɐj			hɔ:k	ka:#		1042 (1042)
Law & Lee	#k		eŋ$ts			ɐj$h	ɔ:k$k		a:#	1801
diphones	#ke	eŋ	ŋ$ts	tsɐj	ɐj	j$hɔ:	ɔ:k	ka:	a:#	1097

Table 3. The string of basic units and exceptional units (underlined) that would be needed to synthesize an utterance of the word 'economist' in each of the three models.

One of these effects is the fusion of coda [p] with initial [h] in words such as 集合 `zaap6 hap6`. In fluent connected speech, such effects can be extreme. Consonants can be reduced or deleted, with the abutting vowels fused together, as in the pronunciation [jy:n^{21}la:^{212}] for the phrase 原來係 `jyun4loi4 hai6` 'was' (with the verb cliticized onto the preceding tense adverb) as in Figure 2. Eventually, we plan to use the larger Segmentation Corpus to look for likely targets of fusion. For now, however, we focus on positional effects that can be modeled just by recording word lists. Figure 2 illustrates one such effect. The final syllable in this utterance is sustained, to be longer than the other four syllables combined. It also bears two extra tone targets for the HL% boundary tone. (See Wong, Chan & Beckman, in press, for a discussion of these utterance-level effects.) Phrase-final lengthening is not usually so extreme in read speech, and the inventory of boundary tones is more limited. However, there will be sufficient final lengthening to warrant the recording of separate units for (the rhymes of) final syllables. These two positional effects increase the inventory of units, albeit in different ways depending on the choice of "basic" unit.

4 Counting different unit types

Table 3 illustrates three synthesis models using the word 經濟學家 `ging1zai3hok6gaa1` 'economist'. The last column in Table 3 lists the theoretically possible number of basic units. The first model concatenates syllables. If each onset could be combined with each rhyme, there would be 1042 syllable types. A second set of syllables can be recorded to capture final lengthening. However, there is no comparably obvious way to capture the cross-syllabic effects with this concatenative unit. The second model uses cross-syllabic units which combine the rhyme of one syllable with the following initial.

This automatically captures the sandhi effects. The model also captures final lengthening, because separate units are recorded for onsets with no preceding rhyme and rhymes with no following onset. With 54 final rhymes, 1728 combinations of rhyme followed by medial onset, and 19 initial onsets, there are 1801 theoretically possible units. The last model is our diphone model, which differentiates codas from onset consonants. That is, the rhyme `aak$` is distinct from the cross-syllabic diphone `aa$k`. This model has the advantage over Law & Lee's cross-syllable final-initial combination model in that spectral continuity between the initial and rhyme is captured in the CV diphones (such as `#gi` and `zai`). Similarly, the diphones capture the dependency between the quality of the [h] and that of the following vowel (i.e., one records separate cross-syllable diphones for `i$ho`, `i$hi`, `i$haa`, and so on). However, the number of theoretically possible units is smaller, because we do not record consonant sequences that abut silence with silence — e.g., `aak$` can be combined directly with `$ka` or `$ta`, so no cross-syllabic units need be recorded for `k$k` and `k$t`.

Note that none of these counts takes tone into consideration. However, since every syllable bears a (full) tone, and since tones are rarely deleted in running speech, recording different units for rhymes with different tones should improve naturalness, particularly for cases where voice quality is part of the tonal specification (as suggested by the contour for tone 4 in Figure 1). Naturalness may also require different cross-syllabic units for different tone sequences when sonorant segments abut at syllable edges (so as to insure tonal continuity).

Of course, when tone specification is taken into account, the number of necessary units grows substantially. For example, there are 12,120 distinct syllables, and even more units in the other two models. However, when we count only those types that are attested in the words of the Segmentation Corpus, there are many fewer units. For example, the total number of attested units taking tone into account in the diphone model is 2292. If each diphone were recorded in a disyllabic carrier word, a Cantonese speaker could speak all of the words to make a new voice in a single recording session. (For comparison, the number of attested diphones ignoring tone is 634.)

Conclusion

We have shown one way of using a segmented database to inform the design of a unit inventory for TTS. We augmented the Segmentation Corpus with transliterations that would let us predict more accurately the pronunciation that a Cantonese speaker adopting a careful speaking style would be likely to produce for a character sequence. Judgements about the phonology of Cantonese, in combination with the augmented wordlist, and the associated word frequency data, can be used to assess the costs and likely benefits of different strategies for unit selection in Cantonese TTS. In particular, we present data indicating the feasibility of a new diphone selection strategy that finesses some of the problems in modelling the interactions between tone and segmental identity. It remains to be demonstrated that this strategy can actually deliver the results which it appears to promise. This is our future work.

Acknowledgements

This work was supported by a grant from the University Grants Committee of Hong Kong to Y. S. Cheung and an SBC/Ameritech Faculty Research Award to C. Brew and M. Beckman.

References

Chan S. D. and Tang Z. X. (1999) *Quantitative Analysis of Lexical Distribution in Different Chinese Communities in the 1990's*. Yuyan Wenzi Yingyong [Applied Linguistics], No.3, 10-18.

Chu M. and Ching P. C. (1997) *A Cantonese synthesizer based on TD-PSOLA method.* Proceedings of the 1997 International Symposium on Multimedia Information Processing. Academia Sinica, Taipei, Taiwan, Dec. 1997.

Law K. M. and Lee Tan (2000) *Using cross-syllable units for Cantonese speech synthesis*. Proceedings of the 2000 International Conference on Spoken Language Processing, Beijing, China, Oct. 2000.

Wong W. Y. P., Chan M. K-M., and Beckman M. E. (in press) *An Autosegmental-Metrical analysis and prosodic conventions for Cantonese*. To appear in S-A. Jun, ed. Prosodic Models and Transcription: Towards Prosodic Typology. Oxford University Press.